THE ESTIMATION OF CONDITIONAL DENSITIES

by

Xiaohong Chen, Oliver Linton, Peter M Robinson[1]
London School of Economics and Political Science

Contents:

The Suntory Centre
Suntory and Toyota International Centres for Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Te.: 020 7955 6698

Discussion paper
No.EM/01/415
May 2001

---

[1] Research supported by a Leverhulme Trust Personal Research Professorship and ESRC Grant R000238212.

Abstract

We discuss a number of issues in the smoothed nonparametric estimation of kernel conditional probability density functions for stationary processes. The kernel conditional density estimate is a ratio of joint and marginal density estimates. We point out the different implications of leading choices of bandwidths in numerator and denominator for the ability of the estimate to integrate to one and to have finite moments. Again bearing in mind different bandwidth possibilities, we discuss asymptotic theory for the estimate: asymptotic bias and variance are calculated under various conditions, an extended discussion of bandwidth choice is included, and a central limit theorem is given.

Keywords: Conditional density estimation; serial dependence; bandwidth choice.
JEL No.: C22

Contact address: Professor Peter M Robinson, Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE,
e-mail: pm.robinson@lse.ac.uk

1 Introduction

A distinctive finding in the asymptotic theory of smoothed nonparametric estimation of probability densities, conditional densities and regression functions is that the same first-order asymptotic properties can hold for stationary weakly dependent observations, recorded at fixed, equally-spaced, times, as for independent ones. In particular, precisely the same multivariate central limit theorem holds for a vector of estimates of the function at finitely many fixed points. The limiting variance matrix is diagonal, which is unsurprising in case of independent observations (as is seen on considering the case of kernel estimates with a finite-support kernel) but less so in the dependent case. The fact that asymptotic variances are unaffected by weak dependence contrasts with the typical experience in parametric estimation. Intuitively, it is due to the local character of the estimate, which depends principally on only a vanishing fraction of the observations; these will tend to be widely separated in time and thus be virtually independent.

The first such central limit results, for kernel estimates computed from Markov sequences, were obtained by Roussas (1967, 1969), who considered a univariate central limit theorem, and by Rosenblatt (1970, 1971), who went on to establish the asymptotic independence property. The same type of result was then established for α-mixing processes by Robinson (1983), Rosenblatt (1985), then for a variety of other mixing processes by various authors, and for linear processes by Chanda (1983). Serial dependence is, however, liable to affect the goodness of the multivariate normal approximation in finite samples. Analytic results of Robinson (1986) in a simple case suggested that it can significantly inflate variance in case of negative as well as positive dependence, so that a larger bandwidth might be used under dependence in order to achieve comparable precision to that available for a corresponding independent sequence. Under long range dependence, even first-order asymptotic properties are considerably affected; estimates may or may not be asymptotically normal, and, if they are, the limiting variance matrix of estimates at distinct fixed points can have unit rank (see e.g. Robinson, 1991, Csörgo and Mielniczuk, 1995).

One of the notable features of the early work of Roussas (1967, 1969) was its discussion of conditional probability density estimation for Markov processes. Indeed, notwithstanding the early results of Parzen (1962) on univariate density estimates, and Cacoullos (1966) on multivariate ones, we know of no earlier work which explicitly considered smoothed nonparametric conditional density estimation even in the case of independent observations, though Rosenblatt (1969) is a roughly contemporary reference in the latter case. Roussas (1969) considered the use of conditional density estimates in estimating the transition distribution function of a Markov process and its quantiles (see also Yakowitz, 1978). There have been a number of subsequent references on conditional density and distribution function estimation under independence and various forms of dependence, but there are some issues that seem worth of further discussion, especially in view of the potential value of the topic in areas of current interest, such as in conditional quantile semiparametric estimation (see e.g. Lee, 1996) or in studying the dynamics of stochastic volatility, where the ability to avoid moment

conditions on the underlying process is an advantage, in view of the long-tailedness of much financial data.

In the following section, we introduce the kernel conditional density estimate (and consequent estimates of conditional distribution function and quantiles) and consider its precise implementation, with reference to the relative choices of bandwidths in the numerator (a bivariate density estimate) and denominator (a marginal density estimate); two such choices have been stressed in the literature, and we point out that one of these ensures that estimate integrates to one and has finite moments, while the other does not. Section 3 provides asymptotic theory under mixing conditions, stressing again the same two leading bandwidth choices; in particular we discuss asymptototic bias and variance, optimal (minimum mean squared error and plug-in) bandwidth choice and central limit theory.

2 Kernel Conditional Density Estimates

Let us consider a bivariate random variable (Y, Z), having absolutely continuous distribution function with respect to Lebesgue measure. Then the conditional probability density function of Y given Z is

$$f(y \mid z) = \frac{f(y,z)}{f(z)},$$

where $f(.,.)$ is the joint probability density functions of (Y, Z) and $f(.)$ is the density function of Z, assumed positive at the point z. Given observations (Y_i, Z_i), $i = 1, \ldots, n$, we estimate $f(y \mid z)$ by

$$\widehat{f}_{abc}(y \mid z) = \frac{\widehat{f}_{ab}(y,z)}{\widehat{f}_c(z)}, \tag{2.1}$$

where

$$\widehat{f}_{ab}(y,z) = \frac{1}{nab}\sum_{i=1}^{n} K\left(\frac{y-Y_i}{a}\right) K\left(\frac{z-Z_i}{b}\right), \tag{2.2}$$

$$\widehat{f}_c(z) = \frac{1}{nc}\sum_{i=1}^{n} K\left(\frac{z-Z_i}{c}\right), \tag{2.3}$$

where K is a bounded kernel function, integrating to one, and $a = a_n$, $b = b_n$ and $c = c_n$ are positive bandwidth sequences, which decay to zero as $n \to \infty$. For example, we have $Y_i = X_{i+1}$, $Z_i = X_i$, when the univariate sequence $X_1, \ldots, X_{n+1}$ is observed; then the conditional or predictive density estimate of Roussas (1967) is a special case of (2.1). For general bivariate, possibly independent, observations, the estimate of Rosenblatt (1969) is a special case of (2.1). Given (2.1) the conditional distribution function can be estimated by, for example,

$$\widehat{F}_{abc}(y \mid z) = \int_{-\infty}^{y} \widehat{f}_{abc}(x \mid z)dx$$

and, thence, conditional quantiles by $\widehat{\zeta}_{abc}(p,z)$, satisfying

$$\widehat{F}_{abc}\left(\widehat{\zeta}_{abc}(p,z) \mid z\right) = p, \quad 0 < p < 1.$$

Our unusual stress on the bandwidths a, b, c in the notation $\widehat{f}_c$, $\widehat{f}_{ab}$ and $\widehat{f}_{abc}$ is deliberate, as the literature employs different relative choices of a, b and c, and there are interesting issues pertaining to their choice. The early estimate of Roussas (1967, 1969) took

$$a = b = c^{\frac{1}{2}}. \tag{2.4}$$

This makes sense in that the numerator and denominator of (2.1) are then estimated with comparable precision, a property which we take for granted in another form of ratio kernel estimate, the Nadaraya-Watson regression estimate. The construction (2.4) has also been used by Robinson (1983), Rosenblatt (1985).

Another leading choice of a, b, c is

$$a = b = c. \tag{2.5}$$

Now the denominator $\widehat{f}_c$ is estimated more precisely than the numerator $\widehat{f}_{ab}$. However, (2.5) has two desirable properties not shared by (2.4).

Property 1: $\widehat{F}_{abc}(\infty \mid z) = 1$, for all z.

Property 2: If K is nonnegative, $\widehat{f}_{aaa}(y \mid z)$ is bounded, and thus has finite moments of all orders.

Under (2.5), Property 1 is established by integrating (2.1) over y, and Property 2 by the inequality

$$\left|\widehat{f}_{aaa}(y \mid z)\right| \leq \frac{1}{na}\sup_u K(u)\frac{\sum_{i=1}^{n} K\left(\frac{z-Z_i}{a}\right)}{\sum_{i=1}^{n} K\left(\frac{z-Z_i}{a}\right)} \leq \frac{1}{na}\sup_u K(u).$$

(We have adopted the convention $0/0 = 1$.) On the other hand, under (2.4)

$$\widehat{F}_{abc}(\infty \mid z) = \frac{\widehat{f}_a(z)}{\widehat{f}_{a^2}(z)},$$

which will converge in probability to one under reasonable conditions, but, for $a > a^2$, i.e. $a < 1$ (as is true for large enough n) is greater than one when K is nonnegative and monotonic, while $\widehat{f}_{aaa^2}(y \mid z)$ need not be bounded, for example $\widehat{f}_{a^2}(z)$ can be zero when $\widehat{f}_{aa}(y,z)$ is non-zero, and need not necessarily have moments. The failure of Property 1 has unpleasant implications for $\widehat{\zeta}(p,z)$, and the failure of Property 2 may be associated with unstable behaviour. The construction (2.5) was explored by Rosenblatt (1969), Masry (1989), Samanta (1989), Roussas (1991a), with a recursive version in Roussas (1991b).

Rosenblatt (1985) considered

$$a = b = o(c^{\frac{1}{2}}), \quad \text{as } n \to \infty, \tag{2.6}$$

which includes (2.5) as a special case, showing that the asymptotic variance in the limiting normal distribution of $n^{\frac{1}{2}}a\left\{\widehat{f}_{aac}(y \mid z) - f(y \mid z)\right\}$ is less under (2.6) than that of $n^{\frac{1}{2}}a\left\{\widehat{f}_{aaa^2}(y \mid z) - f(y \mid z)\right\}$ pertaining to (2.4) (because the asymptotic variance of the denominator density estimate only contributes in the second case). On the other hand, these results entail conditions which ensure that the bias is of sufficiently small order to permit the centring at $f(y \mid z)$. Calculation of leading terms of the bias of the conditional density estimates indicates that the contribution of the denominator will be of smaller order under (2.4) than under (2.6), and thus (2.5).

We shall focus in the sequel on the construction (2.1), (2.3) with Roussas'(1967,1969) transition function case in which $(Y_i, Z_i) = (X_i, X_{i+1})$,though extensions to more general Y_i, Z_i are readily deduced. Further, we shall fix $a = b$ throughout, though it is worth mentioning that generalizations could be of interest. First, while the choice $a = b$ is natural in case $Y_i = X_{i+1}$, $Z_i = X_i$, it need not be for genuine bivariate observations Y_i, Z_i, which can have different scales. Notice here that so long as $b = c$, Properties 1 and 2 will continue to hold. Second, the product kernel in (2.2) could be replaced by a more general bivariate kernel L. Defining A to be a positive definite 2×2 bandwidth matrix we have

$$\widehat{f}_A(y, z) = \frac{1}{n\,|A|} \sum_{i=1}^{n} L\left(A^{-\frac{1}{2}} \begin{pmatrix} y - Y_i \\ z - Z_i \end{pmatrix}\right), \tag{2.7}$$

$$\widehat{f}_{Ac}(y \mid z) = \frac{\widehat{f}_A(y, z)}{\widehat{f}_c(z)}.$$

For one class of L, we can write (2.7) as

$$\widehat{f}_A(y, z) = \frac{1}{n\,|A|} \sum_{i=1}^{n} M\left(\frac{y - Y_i - \frac{a_{12}}{a_{22}}(z - Z_i)}{a_{11} - a_{12}^2/a_{22}}\right) K\left(\frac{z - Z_i}{a_{22}}\right),$$

where a_{ij} is the (i, j)th element of A (for example, when L is a bivariate normal density, we have $M = K$). Then if M is bounded and $a_{22} = a$, Properties 1 and 2 hold. Extensions to higher dimensions, and to alternative forms of nonparametric estimation, such as local polynomials, can also be considered.

3 Asymptotic Theory of Conditional Density Estimates and Bandwidth Choice

We first provide a list of assumptions useful in asymptotic theory for the kernel conditional density estimate given by (2.1)-(2.3) with $(Y_i, Z_i) = (X_i, X_{i+1})$. We shall only consider short range dependent X_i. As the discussion in the Introduction suggests, the literature indicates that a variety of

assumptions on X_i, in particular of Markov, mixing or linear type, will lead to the same type of first-order asymptotic theory. Moreover various trade-offs are possible, involving these and the other types of assumption involved, which concern marginal properties and the kernel and bandwidths. Here we gather assumptions most similar to those imposed by Robinson (1983) and Roussas (1991a).

Assumption P: (i) The sequence of real-valued random variable $\{X_i, i \geq 1\}$ is strictly stationary, and the marginal distribution function of each X_i is absolutely continuous (with respect to Lebesgue measure), with continuous positive density function $f(\cdot)$.

(ii) For each $i \geq 2$, the joint probability distribution distribution function of (X_1, X_i) is absolutely continuous (with respect to Lebesgue measure), with continuous joint density.

Assumption $\mathbf{M}_\alpha$: (i) The sequence $\{X_i, i \geq 1\}$ is strong mixing (α-mixing), i.e.,

$$\alpha_j = \sup_t \sup \left\{ |P(A \cap B) - P(A)P(B)| : A \in \mathcal{M}_{-\infty}^i, B \in \mathcal{M}_{i+j}^\infty \right\} \to 0 \text{ as } j \to \infty.$$

(ii) The $\alpha-$ mixing coefficients $\alpha(j)$ satisfy $\sum_{j=N}^\infty \alpha_j = o(N^{-1})$ as $N \to \infty$.

Assumption $\mathbf{M}_\beta$: (i) The sequence $\{X_i, i \geq 1\}$ is a (Harris) positive recurrent Markov process with a single ergodic set and no cyclically moving subsets, and is initialized from its invariant distribution.

(ii) The sequence $\{X_i, i \geq 1\}$, is absolutely regular (β-mixing), i.e. $\sum_{j=1}^\infty \beta(j) < \infty$, where

$$\beta(j) = \int \sup_{|g| \leq 1} \left| \int g(y)[f_j(y|z) - f(y)]dy \right| f(z)dz \to 0 \text{ as } j \to \infty, \tag{3.1}$$

such that $f_j(\cdot|x)$ is the conditional density of X_{1+j} on $X_1 = x$;the β-mixing coefficients $\beta(j)$ satisfy $\sum_{j=1}^\infty \beta(j) < \infty$.

Assumption K: the kernel $K(\cdot)$ is a probability density function defined on the real-line such that: **(i)** K is bounded and symmetric around zero; **(ii)** $|u|\, K(u) \to 0$ as $|u| \to \infty$; **(iii)** $\int u^2 K(u)du < \infty$.

Assumption B: $\max(a, c) \to 0$, $n \times \max(a^2, c) \to \infty$ as $n \to \infty$.

Assumption D: f is positive and has a continuous and bounded second derivative at z; $f(\cdot, \cdot)$ has continuous second partial derivatives at (y, z).

Remarks: (1) Assumptions P, M and D describe the dependence and marginal properties of X_i. P and M_β(i) are the same as Roussas's (1991a) A1(i) (ii). We replace Roussas's assumption A1(iii) on $\rho - mixing$ with $\rho(j) = O(j^{-\nu})$ for $\nu > 1$ by $\beta - mixing$ with $\sum_{j=1}^\infty \beta(j) < \infty$ in M_β(ii). By Davydov (1973), for a stationary homogeneous Markov process, the definition of $\beta - mixing$ (3.1) is equivalent to:

$$\beta(j) \equiv \int \left| P^j(x, \cdot) - \nu(\cdot) \right|_{var} \nu(dx) \to 0 \text{ as } j \to \infty$$

where $|\cdot|_{var}$ denotes the total variation norm of a signed measure; $\nu(\cdot)$ is the stationary invariant measure; and $P^j(x, A) \equiv \Pr(X_{1+j} \in A \mid X_1 = x)$, the j-step transition probability kernel. Under Assumptions P and M_β(i), this definition is equivalent to that in Assumption M_β(ii). Moreover,

Assumption M_β(i) implies that $\{X_t\}$ is stationary $\beta-mixing$, and hence $\alpha-mixing$. In addition to M_β(i), Roussas (1967, 1969) assumed condition (D_0) which is equivalent to uniform ($\phi-$) mixing with exponential decay. Roussas (1991a) relaxed the condition (D_0) to $\rho-mixing$ with decay rate $O(j^{-\nu})$ for $\nu > 1$. However, by Bradley (1986), for a stationary Markov process, either the $\phi-mixing$ or $\rho-mixing$ coefficient is identically one for all time lags or it decays exponentially. In general $\alpha-mixing$ allows for more temporal dependence than $\beta-mixing$ and $\rho-mixing$, which is why authors such as Robinson (1983), Rosenblatt (1985), Masry (1989) and Bosq (1996) have assumed $\alpha-mixing$. However under M_β(i), $\alpha-mixing$ has the same decay rate as $\beta-mixing$, see e.g. Rosenblatt (1985), Davydov (1973) and Bradley (1986). Another advantage of the $\beta-mixing$ assumption is that it automatically implies an assumption often used in asymptotic theory for nonparametric density estimation based on other types of mixing processes:

$$|f_j(y,z) - f(y)f(z)| \leq C < \infty, \text{ for all } j \geq 1, x, y \in \Re$$

where $f_j(\cdot,\cdot)$ is the joint density of (X_1, X_{1+j}); see e.g. Roussas(1991a, Assumption A5(iii)), Masry (1989) and Bosq (1996). For these reasons one might prefer M_β(ii) over other mixing assumptions, given assumptions P and M_β(i). However, the Markov assumption M_β(i) is not really needed for the results in this paper, though it does partially motivate interest in $f(y\,|z)$. We shall only assume that the stationary sequence $\{X_t\}$ satisfies P and M_α or M_β. Notice that even though M_β(i) implies M_α(i), M_β(ii) may not imply M_α(ii).

(2) Assumption K is the same as Assumptions A2 (i),(ii),(iii) of Roussas(1991a), and imposes no serious practical restriction on the kernel.

(3) Assumption B is a minimal restriction on the bandwidth numbers for the central limit theorem to hold.

(4) Assumption D is the same as Assumption A5 (i)(ii) of Roussas(1991a), and is again standard, though it would be possible to obtain results under milder smoothness assumptions, or indeed under stronger ones if Assumption K were relaxed to permit use of higher order kernels; in particular this would affect the order of magnitude of the asymptotic bias, see Lemma i below.

We first discuss the asymptotic bias $AB(y,z)$, by which we mean the leading terms in the deviation of the ratio of expectations of numerator and denominator of $\widehat{f}_{aac}(y\,|z)$ from $f(y\,|z)$;further to our remarks of the previous section, there is no presumption that the expectation of $\widehat{f}_{aac}(y\,|z)$exists. Define

$$B_1(y,z) = \frac{\int u^2 K(u)du}{2f(z)}\{\frac{\partial^2 f(y,z)}{\partial y^2} + \frac{\partial^2 f(y,z)}{\partial z^2}\},$$

$$B_2(y,z) = \frac{\int u^2 K(u)du}{2f(z)} f(y\,|z)\ \frac{d^2 f(z)}{dz^2}.$$

Lemma 1 *Under Assumptions P, K, D and $a, c \to 0$ as $n \to \infty$, we have:*

$$AB(y,z) = a^2 B_1(y,z) - c^2 B_2(y,z) + o\left(\max\{a^2, c^2\}\right).$$

Hence for case (2.4) $a = b = c^{\frac{1}{2}}$*, we have:*

$$AB(x,y) = a^2 B_1(y,z) + o(a^2), \tag{3.2}$$

and for case (2.5) $a = b = c$*, we have:*

$$AB(x,y) = a^2\{B_1(y,z) - B_2(y,z)\} + o(a^2). \tag{3.3}$$

This is proved by a standard Taylor series argument, as in Rosenblatt (1969, 1985), Robinson (1983).

We next consider $AV(y,z)$, the asymptotic variance of $\widehat{f}_{aac}(y\,|z)$, where this refers to the variance in the limit distribution, and makes no presumption that $\widehat{f}_{aac}(y\,|z)$ has finite variance. Define

$$V_1(y,z) = \int K(u)^2 du \frac{f(y\,|z)^2}{f(z)},$$

$$V_2(y,z) \;\; = \;\; \{\int K(u)^2 du\}^2 \frac{f(y\,|z)}{f(z)},$$

$$V_3(y,z,a,c) = 2\int K(u)K(\frac{au}{c})du\frac{f(y\,|z)^2}{f(z)}.$$

Lemma 2 *Under Assumptions P,* M_α *or* M_β*, K, B and D, we have:*

$$AV(y,z) = \frac{V_1(y,z)}{nc} + \frac{V_2(y,z)}{na^2} - \frac{V_3(y,z,a,c)}{nc} + o\left(\frac{1}{n \times \min\{a^2, c\}}\right).$$

If, further, either

$$a = O(c^{1/2}),\;\; c = o(a), \tag{3.4}$$

or

$$a = O(c), \tag{3.5}$$

then

$$AV(y,z) = \frac{V_1(y,z)}{nc} + \frac{V_2(y,z)}{na^2} + o\left(\frac{1}{n \times \min\{a^2, c\}}\right). \tag{3.6}$$

Hence for case (2.4), $a = b = c^{\frac{1}{2}}$*, we have*

$$AV(y,z) = \frac{V_1(y,z)}{na^2} + \frac{V_2(y,z)}{na^2} + o\left(\frac{1}{na^2}\right), \tag{3.7}$$

while for case (2.5), $a = b = c$*, we have*

$$AV(y,z) = \frac{V_2(y,z)}{na^2} + o\left(\frac{1}{na^2}\right). \tag{3.8}$$

This is proved via a standard linearization argument (see Roussas (1967,1969), Rosenblatt (1969), along with the use of the α-mixing and other assumptions to show that the outcome is identical to that when the X_i are independent observations, as in Robinson (1983). The fact that $V_3(y,z,a,c)$ is absent in (3.6) when we impose (3.4) or (3.5) follows from a dominated convergence argument. Note that (2.4) implies (3.4) and (2.5) implies (3.5).

Comparing with case (2.4), the variance is always smaller larger in case (2.4) because the contribution to variance from the denominator is negligible. AS indicated by Lemma 1, however, the bias may be more or less. For example, suppose that the $X_{\ i}$ are independent standard normal random variables. Then in case (2.5), the variance is proportional to $\phi(y)/\phi(z)$, where ϕ is the standard normal density function, while the bias is proportional to $(y^2-1)\phi(y)$. In case (2.4), the variance is proportional to $[1+\phi(y)]\phi(y)/\phi(z)$, while the bias is proportional to $(y^2+z^2-2)\phi(y)$. Note that in case (2.4) the bias increases as z increases and at a quadratic rate, while in case (2.5) the bias is bounded in y and does not depend on z. Therefore, it is easy to find cases where (2.5) is better than (2.4) and vice-versa. The conclusion is that there is no uniform ranking in terms of asymptotic mean squared errors.

It is possible to apply Lemmas 1 and 2 in bandwidth selection. Denote by $AMSE(y,z) = AV(y,z) + AB(y,z)^2$ the asymptotic mean squared error of $\widehat{f}_{aac}(y\,|z)$. Then in case (2.4) $a = b = c^{1/2}$,we have from (3.2) and (3.7)

$$AMSE(y,z) = \frac{V_1(y,z)+V_2(y,z)}{na^2} + a^4 B_1(y,z)^2.$$

Thus by elementary calculus the a minimizing $AMSE(y,z)$ in this case is

$$a(y,z) = \left\{ \frac{V_1(y,z)+V_2(y,z)}{2nB_1(y,z)^2} \right\}^{1/6}.$$

On the other hand in case (2.5) $a = b = c$,we have from (3.3) and (3.8)

$$AMSE(y,z) = \frac{V_2(y,z)}{na^2} + a^4\{B_1(y,z) - B_2(y,z)\}^2,$$

so that the a minimizing $AMSE(y,z)$ is

$$a(y,z) = \left\{ \frac{V_2(y,z)}{2n\left[B_1(y,z) - B_2(y,z)\right]^2} \right\}^{1/6}.$$

Note that the optimal bandwidth is in both cases of order $n^{-1/6}$, unlike the $n^{-1/5}$ rate familiar from univariate density estimation. In practice the $B_i(y,z)$ and $V_i(y,z)$ might be replaced by consistent estimates in order to construct feasible, approximately optimal bandwidths. Our existing estimates $\widehat{f}_{aac}(y\,|z)$ and $\widehat{f}_c$ might be used for this purpose, along with kernel estimates of $d^2f(z)/dz^2$, $\partial^2 f(y,z)/\partial y^2$ and $\partial^2 f(y,z)/\partial z^2$.

In practice it may be desired to estimate $f(y\,|z)$over a range of y and z values, and to use a global bandwidth. Considering first a bandwidth that is global with respect to y but pointwise with respect to z, we might consider the asymptotic integrated mean squared error

$$AIMSE(z) = \int \{AV(y,z) + AB(y,z)^2\}dy.$$

In case (2.4) $a = b = c^{1/2}$ this gives the optimal bandwidth

$$a(z) = \left\{ \frac{\int [V_1(y,z) + V_2(y,z)]\, dy}{2n \int B_1(y,z)^2 dy} \right\}^{1/6}. \tag{3.9}$$

In case (2.5) $a = b = c$ it gives instead

$$a(z) = \left\{ \frac{\int V_2(y,z)dy}{2n \int [B_1(y,z) - B_2(y,z)]^2\, dy} \right\}^{1/6}. \tag{3.10}$$

Notice that these formulae can be slightly simplified due to the identity

$$\int V_2(y,z)dy = \{\int K(u)^2 du\}^2/nf(z).$$

The remaining integrals might be approximated by summations, possibly over data points.

Deriving optimal bandwidths that are global with respect to both y and z is problematic as $AMSE(y,z)$ need not be integrable with respect to z. One solution is to use the above formulae evaluated with, say, z replaced by the sample mean or median of X_i. The computations are still somewhat onerous, however, while if smoothed nonparametric estimates are used the precision of the approximately optimal bandwidths is in doubt.

A solution is to employ instead a 'pilot' parametric distribution in (3.9) and (3.10), in the spirit of Silverman(1986). He considered univariate density estimation from independent observations; our case of conditional density estimation from observations that are likely dependent is more complex. Suppose we proceed as if X_i is a stationary Gaussian process with mean μ,variance σ^2 and lag-one autocorrelation ρ (Silverman (1986) took $\rho = 0$ in his univariate density cae). Then the density of X_i would be given by

$$f(z) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{ -\frac{(z-\mu)^2}{2\sigma^2} \right\},$$

while the joint density of (X_i, X_{i+1}) would be given by

$$f(y,z) = \frac{1}{2\pi\sigma^2(1-\rho^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2(1-\rho^2)} \left[(y-\mu)^2 - 2\rho(y-\mu)(z-\mu) + (z-\mu)^2\right]\}.$$

It follows from routine calculation that

$$\begin{aligned}
B_1(y,\mu) &= \frac{\int u^2K(u)du}{(4\pi)^{1/2}\sigma^3(1-\rho^2)^{3/2}}\left\{\frac{(1+\rho^2)(y-\mu)^2}{(1-\rho^2)\sigma^2}\right\}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2(1-\rho^2)}\right\},\\
B_2(y,\mu) &= -\frac{\int u^2K(u)du}{(4\pi)^{1/2}\sigma^3(1-\rho^2)^{1/2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2(1-\rho^2)}\right\},\\
V_1(y,\mu) &= \frac{\int K(u)^2du}{(2\pi)^{1/2}\sigma(1-\rho^2)}\exp\left\{-\frac{(y-\mu)^2}{\sigma^2(1-\rho^2)}\right\},\\
V_2(y,\mu) &= \{\int K(u)^2du\}^2(2\pi)^{1/2}\sigma
\end{aligned}$$

Thence, after lengthy calculation we deduce from (3.9) that under (2.4) $a = b = c^{1/2}$,an optimal bandwidth (at $z = \mu$) based on the Gaussian prescription is

$$\widetilde{a}(\mu) = \left\{\frac{8(4\pi)^{1/2}\sigma^5(1-\rho^2)^{5/2}\left[(2\pi)^{1/2}(\int K(u)^2du)^2 + \frac{\int K(u)^2}{\sqrt{2}(1-\rho^2)^{1/2}}\right]}{n(3\rho^4-2\rho^2+11)\left[\int u^2K(u)du\right]^2},\right\}^{1/6} \tag{3.11}$$

while under (2.5) $a = b = c$ we have from (3.10)

$$\widetilde{a}(\mu) = \sigma\left\{\frac{16\pi\sqrt{2}(1-\rho^2)^{5/2}\left[\int K(u)^2du\right]^2}{n(15\rho^4-50\rho^2+39)\left[\int u^2K(u)du\right]^2}\right\}^{1/6}. \tag{3.12}$$

Of course the intention is that these be used without the Gaussianity assumption, in which case they are not optimal, though they do have the optimal raste of convergence, adapt naturally to scale at least, and may hopefully be useful when the actual process is not close to being Gaussian. Notice that both formulae (3.11) and (3.12) are invariant with respect to μ, but depend on the unknown σ and ρ, for which we may insert the sample standard deviation and lag-1 sample autocorrelation

$$\begin{aligned}
\widehat{\sigma} &= \{\frac{1}{n}\sum_{i=1}^{n+1}(X_i-\overline{X})^2\}^{1/2},\\
\widehat{\rho} &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(X_{i+1}-\overline{X})}{\widehat{\sigma}^2},
\end{aligned}$$

respectively, where $\overline{X} = (n+1)^{-1}\sum_{i=1}^{n+1} X_i$; due to the $n^{1/2}-$ consistency of $\widehat{\sigma}$ and $\widehat{\rho}$ (under the assumptions listed above along with finite fourth moments of X_i) the consequent plug-in bandwidths should be fairly stable in moderate samples, even though they are not in general approximately optimal. A simplified version, suitable for independent X_i, puts $\rho = 0$ in (3.11) and (3.12).

We next consider central limit theory.

Theorem 3 *Let Assumptions P, M_α or M_β, K, B and D be satisfied, and $a, c > 0$ be such that:*

$$\lim_{n\to\infty}[n \times \min\{a^2, c\} \times \max\{a^4, c^4\}] = 0 \tag{3.13}$$

$$\lim_{n\to\infty} \min(1, \frac{c}{a^2}) + \lim_{n\to\infty} \min(1, \frac{a^2}{c}) > 0 \quad \textit{exists.}$$

Then we have

$$\sqrt{n \times \min\{a^2, c\}} \left(\widehat{f}_{aac}(y\,|z) - f(y\,|z) \right) \Longrightarrow \mathcal{N}\left(0, V(y,z)\right),$$

where

$$V(y,z) = \frac{f(y\,|z)}{f(z)} \int K^2(u)du \,[\, \min\left\{1, \lim_n \frac{c}{a^2}\right\} \int K^2(u)du + \min\left\{1, \lim_n \frac{a^2}{c}\right\} f(y\,|z)\,].$$

Hence, with B simplified to $na_n^2 \to \infty$, *and (3.13) to* $na_n^6 \to 0$, *we have for case (2.4)* $a = b = c^{1/2}$,

$$\sqrt{na^2} \left(\widehat{f}_{aac}(y\,|z) - f(y\,|z) \right) \Longrightarrow \mathcal{N}\left(0, V_1(y,z) + V_2(y,z)\right),$$

and for case (2.5) $a = b = c$,

$$\sqrt{na^2} \left(\widehat{f}_{aac}(y\,|z) - f(y\,|z) \right) \Longrightarrow \mathcal{N}\left(0, V_2(y,z)\right).$$

The theorem is proved by applying Robinson (1983, lemma 7.1) in the $\alpha-$mixing case, and in the $\beta-$mixing case by proceeding similarly but employing results like those of Viennet (1996) . As usual, the asymptotic variances are of the same type as those in case of independent observations in that there is no contribution from 'covariance' terms, though of course in the present instance if we impose the independence by writing $f(y\,|z) = f(y)$, there is some slight simplification in our asymptotic variance formulae. As in Robinson (1983) we can consistently estimate the limiting variances by inserting smoothed nonparametric estimsates of the unknown components, in order to carry out pointwise inferences. These are useful because, as in that reference and Rosenblatt (1970,1971), it is possible to extend the result to a multivariate central limit theorem, indicating asymptotic independence of the $\sqrt{na^2} \left(\widehat{f}_{aac}(y_k\,|z) - f(y_k\,|z) \right)$ across finitely many distinct fixed points $y_1, y_2, ...$

REFERENCES

Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes.* Lecture Notes in Statistics, Springer-Verlag, New York.

Bradley, R. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics* (E. Eberlein and M.S. Taqqu, eds.) Birkhauser, 165-192.

Cacoullos, T. (1966). Estimation of multivariate density. *Ann. Inst. Statist. Math.* 18, 179-189.

Chanda, K.C. (1983). Density estimation for linear processes. *Ann. Inst. Statist. Math.* 35, 439-445.

Csörgo, S. and Mielniczuk, J. (1991). Density estimation under long-range dependence. *Ann. Statist.* 23, 990-999.

Davydov, Y.A. (1973). Mixing conditions for Markov chains. *Theor. Probab. Appl.* 18, 312-328.

Lee, M.-J. (1996). *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models.* Springer-Verlag, New York.

Masry, E. (1989). Nonparametric estimation of conditional probability densities and expectations of stationary processes: strong consistency and rates. *Stoch. Proc. Appl.* 32, 109-127.

Parzen, E. (1962). On the estimation of a probability density and mode. *Ann. Math. Statist.* 35, 1065-1076.

Robinson, P.M. (1983). Nonparametric estimators for time series. *J. Time Series Anal.* 4, 185-207.

Robinson, P.M. (1986). On the consistency and finite-sample properties of nonparametric kernel time series regression, autoregression and density estimators. *Ann. Inst. Statist. Math.* 38, 539-549.

Robinson, P.M. (1991). Nonparametric function estimation for long memory time series. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (W.A. Barnett, J. Powell and G.E. Tauchen, eds.) Cambridge University Press, Cambridge, 437-457.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis II* (P.R. Shnaiah, ed.) Academic Press, New York, 25-31.

Rosenblatt, M. (1970). Density estimates and Markov processes. In *Nonparametric Techniques in Statistical Inference* (M.L. Puri, ed.) Cambridge University Press, Cambridge, 199-210.

Rosenblatt, M. (1971). Curve estimates. *Ann. Math. Statist.* 42, 1815-1842.

Rosenblatt, M. (1985). *Stationary Sequences and Random Fields.* Birkhäuser, Boston.

Roussas, G. (1967). Nonparametric estimation in Markov processes. *Ann. Inst. Statist. Math.* 21, 73-87.

Roussas, G. (1969). Nonparametric estimation of the transition distribution function of a Markov process. *Ann. Math. Statist.* 40, 1386-1400.

Roussas, G. (1991a). Estimation of transition distribution function and its quantiles in Markov processes: strong consistency and asymptotic normality. In *Nonparametric Functional Estimation and Related Topics*, (G. Roussas, ed.) Kluwer, Amsterdam, 443-462.

Roussas, G. (1991b). Recursive estimation of the transition distribution function of a Markov process: asymptotic normality. *Statist. Probab. Lett.* 11, 435-447.

Samanta, M. (1989). Non-parametric estimation of conditional quantiles. *Statist. Probab. Lett.* 7, 407-412.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Viennet, G. (1996). Inequalities for absolutely regular sequences: application to density estimation. *Prob. Theor. Rel. Fields.* 107, 467-492.

Yakowitz, S. (1979). Nonparametric estimation of Markov transition functions. *Am. Statist.* 7, 671-679.